Cancer Breakpoint Hotspots Versus Individual Breakpoints Prediction by Machine Learning Models

Kseniia Cheloshkina[1], Islam Bzhikhatlov[2], and Maria Poptsova[1](✉)

[1] Laboratory of Bioinformatics, Faculty of Computer Science, National Research University
Higher School of Economics, 11 Pokrovsky boulvar, Moscow 101000, Russia
mpoptsova@hse.ru

[2] Faculty of Control Systems and Robotics, ITMO University,
49 Kronverksky Pr., St. Petersburg 197101, Russia

Abstract. Genome rearrangement is a hallmark of all cancers. Cancer breakpoint prediction appeared to be a difficult task, and various machine learning models did not achieve high prediction power. We investigated the power of machine learning models to predict breakpoint hotspots selected with different density thresholds and also compared prediction of hotspots versus individual breakpoints. We found that hotspots are considerably better predicted than individual breakpoints. While choosing a selection criterion, the test ROC AUC only is not enough to choose the best model, the lift of recall and lift of precision should be taken into consideration. Investigation of the lift of recall and lift of precision showed that it is impossible to select one criterion of hotspot selection for all cancer types but there are three to four distinct groups of cancer with similar properties. Overall the presented results point to the necessity to choose different hotspots selection criteria for different types of cancer.

Keywords: Cancer genome rearrangements · Cancer breakpoints · Cancer breakpoint hotspots · Machine learning · Random forest

1 Introduction

Cancer genome rearrangement is a hallmark of all cancers and hundreds of thousands of cancer breakpoints has been documented for different types of cancers [1–3]. Heterogeneity of cancer mutations has been noticed long ago [4] and the accumulated data on cancer genome mutations was termed as cancer genome landscapes [5]. Thousands of cancer full genome data became available to researchers by the International Cancer Genome Consortium (ICGC) [6]. Later, instead of individual cancer genomes a notion of pan cancer has emerged [7, 8] revealing common and individual properties of cancer genome mutations. Recently, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium [9] of the International Cancer Genome Consortium (ICGC) [6] and The Cancer Genome Atlas (TCGA) [10] reported the integrative analysis of more than 2,500

Z. Cai et al. (Eds.): ISBRA 2020, LNBI 12304, pp. 217–228, 2020.
https://doi.org/10.1007/978-3-030-57821-3_19

whole-cancer genomes across 38 tumour types [11]. Apart from [11–13] point mutations, a genome rearrangement with creation of structural elements is often an early event in cancer evolution sometimes preceding point mutation accumulation.

Machine learning methods were successful in finding regularities in studying cancer genome mutations. The most successful machine learning models was shown to be in predicting densities of somatic mutations [14, 15]. In [14] the density of somatic point mutations was predicted by densities of HDNase and histone modifications with determination coefficient of 0.7–0.8. The relative contribution of non-B DNA structures and epigenetic factors in predicting the density of cancer point mutations was studied in [15]. It was shown that taking both groups of factors into account increased prediction power of models.

Despite success in prediction of densities of somatic point mutations, cancer breakpoint prediction models showed low or moderate power [15, 16]. This fact could be explained both by the lack of causal determinants in the models and constrains of the machine learning algorithms.

Previously we showed that the breakpoint density distribution varies across different chromosomes in different cancer types [16]. Also, we showed that determination of hotspot breakpoints depends on a threshold, and the choice of the threshold could vary between different types of cancer and could affect the results of machine learning models. Here we aimed at conducting a systematic study of how breakpoint hotspots density thresholds influence prediction power of machine learning models. We also posed a question whether the prediction power of machine learning models will be different whether we predict individual breakpoints or breakpoint hotspots.

2 Methods

2.1 Data

Data on cancer breakpoints were downloaded from the International Cancer Genome Consortium (ICGC) [6]. The dataset comprises more than 652 000 breakpoints of 2803 samples from more than 40 different types cancers that we grouped in 10 groups of cancer according to tissue types and further refer as cancer types. We cut the genome into non-overlapping windows of 100 KB of length and excluded regions from centromeres, telomeres, blacklisted regions and Y chromosome. Then for each window we estimated breakpoint density as the ratio of the number of breakpoints in the window to the total number of breakpoints in a given chromosome. We used the density metric to designate hotspots, i.e. genomic regions with a relatively high concentration of breakpoints. In the study, we investigate three labeling types of hotspots - 99%, 99.5% and 99.9% percentiles of breakpoint density distribution. Besides, we assigned “individual breakpoints” label to windows containing at least one breakpoint. The proportion of the number of these windows from the total number of windows varied from 2.8% to 90% for different cancer types.

In the study we used the most comprehensive set of predictors, available as of today mostly from next-generation sequencing experiments. The features include genomic regions, TAD boundaries, secondary structures, transcription factor binding sites and a set of epigenetic factors (chromatin accessibility, histone modifications, DNA methylation). These data were collected from The Encode, DNA Punctuation, Non-B DB projects, UCSC Genome Browser. The data were transformed into feature vectors by calculating window coverage of each characteristic.

2.2 Machine Learning Models

After data collection we got 30 datasets that comprise genomic features' coverage and binary target labeling (3 hotspots labeling with different quantile threshold of 99%, 99.5% and 99.9% per each of 10 cancer types).

The hotspot prediction power was evaluated through the train-test splits with stratification by a chromosome with proportion of 70–30, retaining 30% of data for testing. To get a reliable estimate of quality metrics we performed train-test splits 30 times for each dataset because of high class-imbalance (very small ratio of positive examples). For this reason, we also applied the class balancing technique (oversampling) when training a machine learning model. We selected Random Forest as one of the most performing and popular classification algorithm for table data to assess hotspots and breakpoints prediction power. For the model we estimated the best hyperparameters (the number of trees, number of features to grow a tree, minimal number of examples in a terminal node, maximal number of nodes in a tree) by averaging performance metrics among all cancer types.

2.3 Evaluation Metrics

We used several metrics for model evaluation: ROC AUC (Area Under Receiver Operating Curve), precision, recall, lift of precision and lift of recall. As we were dealing with a high class imbalance for each dataset we averaged the results from 30 random train-test splits by taking mean (or median) ROC AUC on the test set and controlling for its standard deviation, which demonstrates how strongly the results depend on the distribution of examples in train and test sets. We calculated recall and precision for different probability percentiles – from 0.5% to 50%. To get an estimate of how well a model performs in comparison with a random choice we used the lift of recall and lift of precision. The lift of recall for a given probability percentile shows how many times the recall of the model (estimated on examples labeled as the positive class according to a probability percentile threshold) is higher than a random choice (it is equal to the recall of the model divided by the probability percentile). Similarly, the lift of precision demonstrates how many times precision for a given probability percentile is higher than a random choice (equal to the precision of the model divided by the proportion of positive examples in a dataset).

3 Results

3.1 Distribution of Test and Train ROC AUC for All Cancer Types by Hotspot Labeling Type

We train Random Forest model on all 30 datasets for hotspot prediction and 10 datasets for individual breakpoint prediction. The distribution of ROC AUC on test set by cancer type and labeling type is given in Fig. 1. It could be seen that for the half of cancer types including blood, brain, breast, pancreatic and skin cancer the higher the hotspot labeling threshold the higher the median of test ROC AUC. For bone, liver, uterus cancer there is no monotonically increasing median test quality but for the highest labeling type this value is higher than for the lowest while for the rest of cancers (ovary and prostate) there is no significant difference between labeling types.

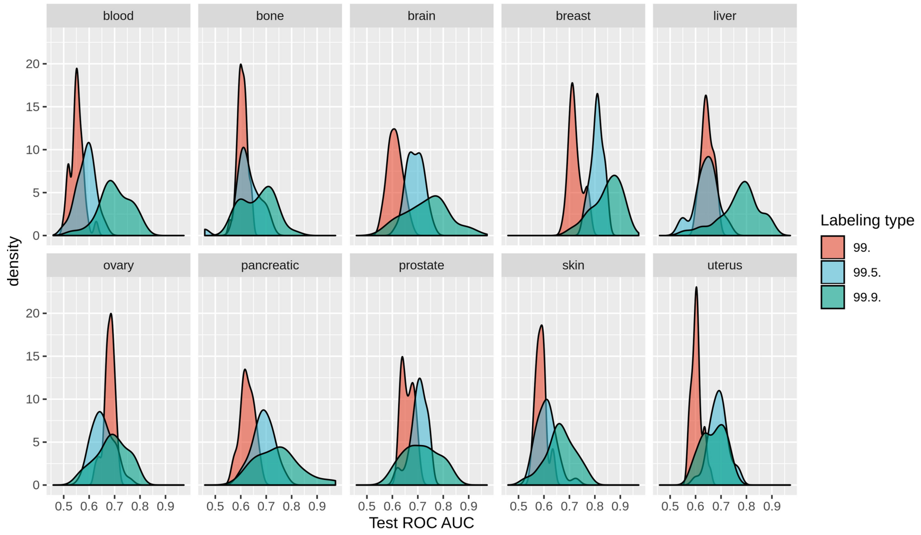

Fig. 1. Distribution of test ROC AUC for all cancer types by hotspots labeling.

Median values of the test ROC AUC by cancer type and labeling type are presented in Fig. 2 and Table 1. The highest quality in terms of considered metric belongs to the breast cancer for all labeling types while the lowest – to the blood and skin cancer for 99% labeling type. The difference greater than 0.10 between the median test ROC AUC for models of the 99% and 99.9% hotspot labeling types is observed for the blood, brain, breast, liver and pancreatic cancer. For the half of the cancer types the highest labeling type implies significantly higher quality according to the median test ROC AUC.

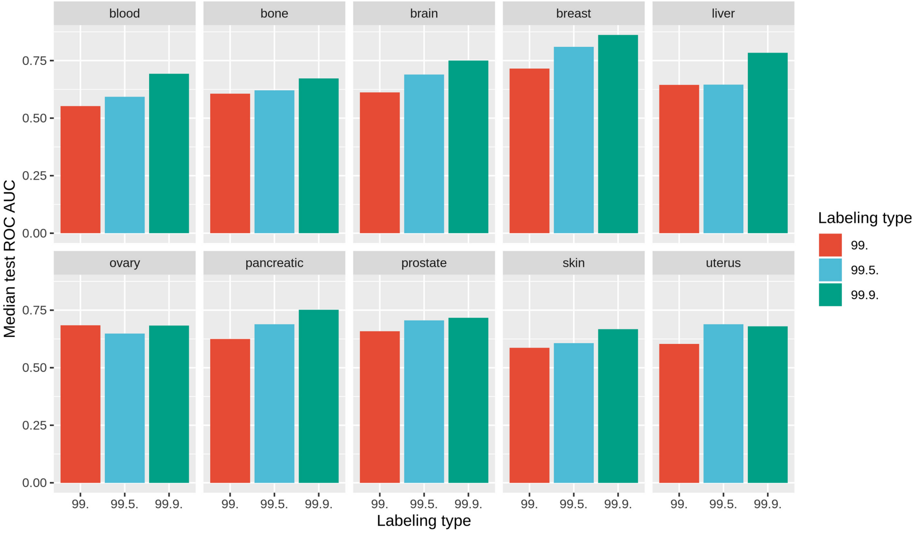

Fig. 2. Median test ROC AUC for each cancer and labeling type.

Table 1. Median test ROC AUC for each cancer and labeling type.

Cancer type	Median test ROC AUC (99%)	Median test ROC AUC (99.5%)	Median test ROC AUC (99.9%)
Blood	0,552	0,593	0,693
Bone	0,606	0,621	0,673
Brain	0,612	0,689	0,75
Breast	0,715	0,81	0,861
Liver	0,645	0,646	0,784
Ovary	0,684	0,648	0,683
Pancreatic	0,625	0,689	0,752
Prostate	0,659	0,706	0,717
Skin	0,587	0,607	0,668
Uterus	0,603	0,689	0,68

On the other hand, as it could be seen in Fig. 3, the more rare hotspots we aim to predict the higher the variance of the test ROC AUC on the test set as well as the difference between the train and test ROC AUC. This could be explained by the fact that for the case of rare hotspots there is small number of positive examples in a dataset and its random permutation between the train and test set leads to different results. Moreover, for all cancer types except for the breast cancer difference between the median train and test ROC AUC for the 99.9% labeling type approaches 0.2 ROC AUC and is 2–3 times higher than for the 99.5% and 99% labeling types. Hence, when selecting the best hotspot labeling type it would be reasonable to choose the 99% or 99.5% labeling type according to the highest median test ROC AUC.

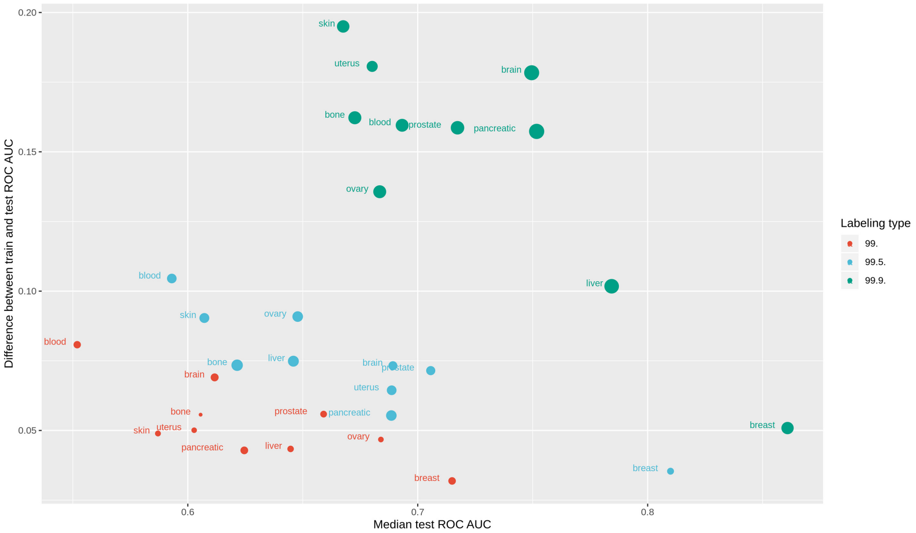

Fig. 3. Dependence of the difference of train and test ROC AUC from test ROC AUC at different labeling types and different standard deviations of test ROC AUC.

3.2 Lift of Precision and Lift of Recall

Next we analyzed the distribution of other quality metrics such as the lift of recall and lift of precision. The results are presented in Fig. 4 and 5 respectively. Here confidence intervals for the mean of these metrics are plotted against different probability quantiles selected as a threshold for model predictions for each cancer and hotspot labeling type. The main conclusion that could be made according to these results is that there is no single labeling type which guarantees the best classification results for all cancer types. However, three groups of cancer types were distinguished: the best labeling type for the blood, brain, liver and pancreatic cancers is 99.9%, for the bone, breast and uterus cancers - 99.5%, for the rest (ovary, prostate, skin cancers) - 99%.

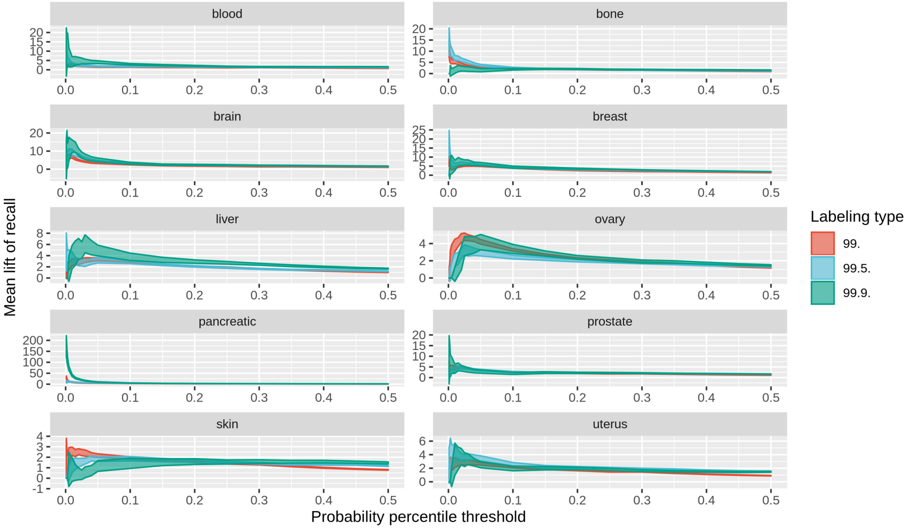

Fig. 4. Dependence of lift of recall from quantile threshold for different aggregation levels (see text for explanation).

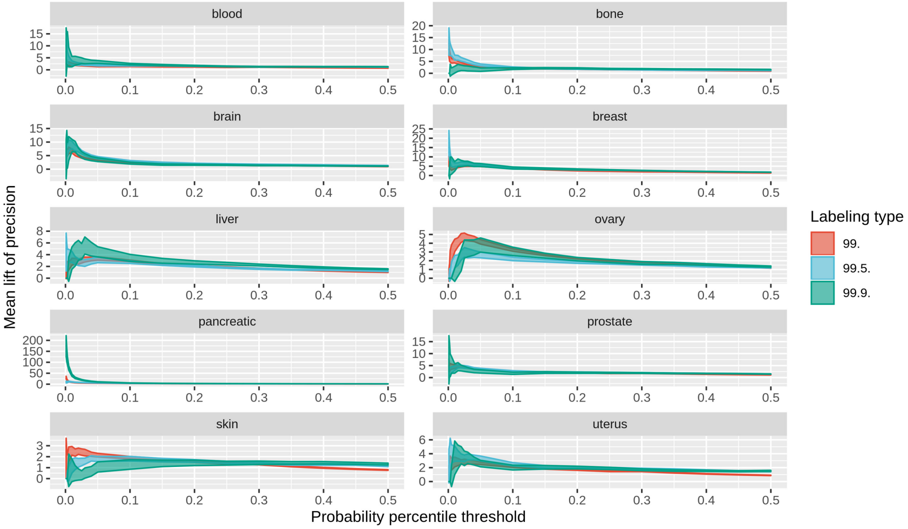

Fig. 5. Dependence of lift of precision from quintile threshold for different aggregation levels (see text for explanation).

When comparing the best labeling type determined with the median test ROC AUC and with the lift of recall/precision, it is the same only for ovary, bone and uterus cancer. As we are mainly interested in selection of minimal number of genome regions with the maximal concentration of hotspots, the final choice of the best labeling type will coincide with the decision according to the lift of recall/precision.

Interestingly, for the breast cancer all three labeling types are almost equally well predicted: they have relatively high lift of recall and differ slightly. Besides, for pancreatic cancer 99.9% labeling type showed significant boost in both lift of precision and lift of recall.

3.3 Prediction of Hotspot Breakpoints Versus Individual Breakpoints

Further, we tested whether recurrent breakpoints could be more effectively recognized by machine learning model than non-recurrent breakpoints and also how well the individual breakpoints are predicted. Distributions of test ROC AUC for hotspots prediction (the best labeling type) and breakpoint prediction tasks are presented in Fig. 6.

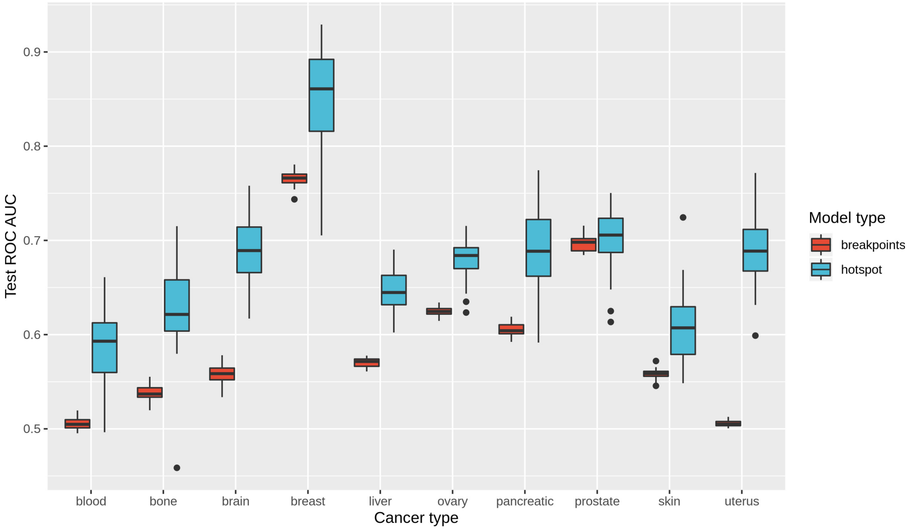

Fig. 6. Distribution of test ROC AUC for the best labeling hotspot profile and all breakpoints prediction.

The observed results can be summarized as follows. Firstly, for the majority of cancer types hotspots are recognized by machine learning models considerably better than individual breakpoints for all cancer types except for the prostate cancer. The quantitative estimate of the difference is given in Fig. 7 and Table 2. The highest ratio of the median test ROC AUC for hotspot prediction model to the median test ROC AUC for breakpoint prediction model is observed for the uterus and brain cancer (1.36 and 1.23 respectively) while for the prostate cancer they are almost equal. For the other cancer types the metric for hotspot model is 9–18% higher than for the breakpoints. Thus, in general, breakpoints are harder to recognize than hotspots using the same genomic features.

Also it could be seen that variance of ROC AUC is significantly lower for breakpoints and this could be a consequence of having a considerably higher number of positive examples for the model.

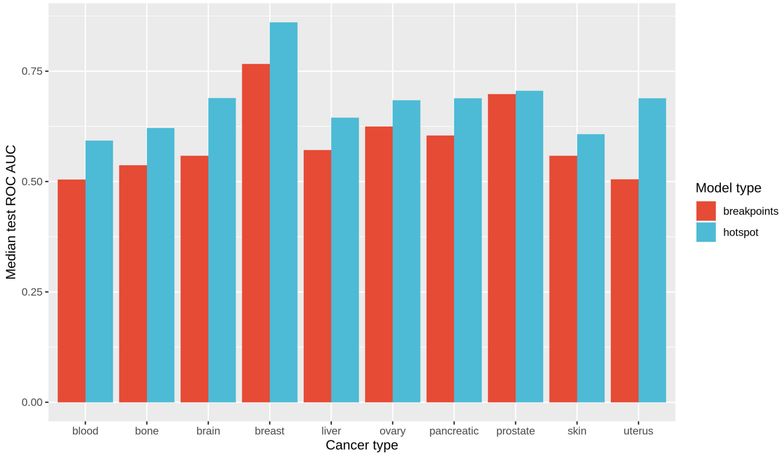

Fig. 7. Median test ROC AUC for best labeling hotspot profile and breakpoints prediction.

Table 2. Median test ROC AUC for best labeling hotspot profile and breakpoints prediction.

Cancer type	Breakpoints median test ROC AUC	Hotspots median test ROC AUC	Ratio of median test ROC AUC for hotspots and breakpoints prediction models
Prostate	0,698	0,706	1,011
Skin	0,559	0,607	1,087
Ovary	0,625	0,684	1,095
Breast	0,766	0,861	1,124
Liver	0,572	0,645	1,128
Pancreatic	0,604	0,689	1,14
Bone	0,537	0,621	1,157
Blood	0,505	0,593	1,175
Brain	0,559	0,689	1,234
Uterus	0,505	0,689	1,363

Secondly, the quality of breakpoint prediction is quite low so that it is a difficult task to predict cancer breakpoints by a machine learning model. For 6 cancer types including skin, liver, bone, blood, brain and uterus cancer the median test ROC AUC does not exceed 0.6. In contrast, the highest value of the metric (0.77) is achieved for breast cancer.

The conclusion is confirmed by the statistics of the lift of recall given in Fig. 8. All cancer types could be divided into 2 groups. For the pancreatic and skin cancer breakpoints are unrecognizable as the lift of recall is very low (almost equal to zero). For the ovary, breast, uterus and prostate the metric hardly achieves 1 for the probability percentile threshold of 0–0.1 so that in these cases breakpoints are predicted as successfully as in the case of a random choice. For the blood, bone, brain and liver cancers there are some probability thresholds for which the lift of recall is higher than 1 with the brain cancer model performing the best. In total, for 6 cancer types the breakpoint prediction model quality does not significantly differ from a random choice and only for 4 cancer types the prediction is slightly better.

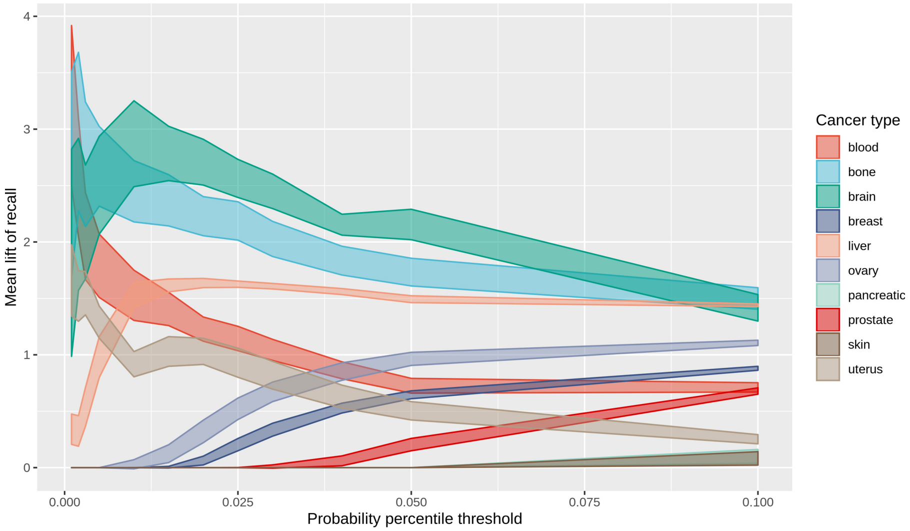

Fig. 8. Lift of recall for breakpoints prediction model

Additionally, it should be noted that for the blood, bone and brain cancer the lift of recall decreases with probability threshold. This could mean that for these cancer types some breakpoints are highly pronounced and could be more easily identified than the rest of breakpoints.

Besides, a set of cancer type models achieving the best performance for the task of breakpoint prediction according to the median ROC AUC (prostate, ovary, breast) differs from a set determined by the lift of recall (blood, bone, brain and liver). This difference outlines the fact that it is very important to choose the right performance metric for a given machine learning task. As the ROC AUC measures a quality of overall examples' ordering produced by the model and the lift of recall measures ordering quality of examples with the highest probabilities, they describe the model performance from different perspectives.

4 Conclusions and Discussion

In this study, using machine learning approach we systematically investigated the effect of different selection criteria for cancer breakpoint predictions for 10 large types of cancer. We built machine-learning models predicting hotspot breakpoints defined by different density thresholds and investigated distributions of the train and test ROC AUC as well as the lift of precision and recall. Almost for all types of cancer the median test ROC AUC increases with an increase of threshold for hotspot selection, though the total quantity of those regions decreases and variance of quality metrics grows up. This fact confirms the fact that the machine learning models better recognize regions with increased density of breakpoint mutations, or regions with recurrent breakpoints, which requires further research. This result could be considered as expected however we empirically found an exception for prostate cancer where median test ROC AUC for hotspots and individual breakpoint do not differ much both being close to 0.70. This suggests that mutagenic processes of individual and recurrent breakpoints in prostate cancer most likely have similar nature. However the effect needs further investigation.

We would like to emphasize that, without actual tests of machine-learning performance on individual breakpoints and breakpoint hotspots, it is not evident, which one of the two will have a higher prediction power. Indeed, breakpoint hotspots are regions enriched with breakpoints, and genomic features of these regions should explain their recurrent formation. On the opposite, rare events are often harder to predict except for the cases when strong predictors of a rare event are available. In the research we posed a question whether the considered genome features identify breakpoints hotspots better than individual breakpoints, and whether individual breakpoints have also distinctive features that influence their formation. The comprehensive analysis of the features is out of scope of the present study is the subject of further systematic research.

The lift of recall and lift of precision signify how many times recall or precision is higher compared to a random choice. Analysis of the distributions of the lift of recall and lift of precision showed that it is impossible to choose one breakpoint density threshold that would lead to the maximum prediction power of models for all types of cancer. Three groups of cancer with similar behavior according to the lift of recall and lift of precision were distinguished. Common properties of cancer breakpoint formation in these three groups of cancer require further investigations.

Selection criteria for the best hotspot labeling threshold based on the median test ROC AUC and the lift of recall and precision coincide only for three types of cancer – ovary, bone, uterus. Moreover, concerning evaluation of prediction power of breakpoints these metrics produce different results. As the ROC AUC and lift of recall measure quality of examples' ordering by the model at different scales (based on all examples and examples with the highest probabilities respectively) we recommend to use the lift of recall and lift of precision metrics to choose the hotspot thresholds.

Comparison of breakpoint predictions and breakpoint hotspots with a chosen selection criterion based on the best machine-learning model showed that the median test AUC is always higher for hotspots rather than for individual breakpoints. We tested additionally several machine learning models such as logistic regression and XGBoost (results are not presented here due to the paper size limitation) and they all show approximately the same relative distribution of ROC AUC, lift of recall and lift of precision across different types of cancer.

Overall the results of our study showed that machine learning model prediction power depends on density threshold for cancer hotspots, and the threshold is different for different types of cancer. Besides, we demonstrated that though individual breakpoints are harder to predict than breakpoint hotspots, individual breakpoints can be predicted to a certain extent, and, moreover, in prostate cancer they are predicted equally well as hotspots. While choosing a selection criterion, the test ROC AUC only is not enough to choose the best model, the lift of recall and lift of precision should be taken into consideration at the level of individual type of cancer.

References

1. Harewood, L., et al.: Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. Genome Biol. **18**, 125 (2017)
2. Nakagawa, H., Fujita, M.: Whole genome sequencing analysis for cancer genomics and precision medicine. Cancer Sci. **109**, 513–522 (2018)
3. Nakagawa, H., Wardell, C.P., Furuta, M., Taniguchi, H., Fujimoto, A.: Cancer whole-genome sequencing: present and future. Oncogene **34**, 5943–5950 (2015)
4. Salk, J.J., Fox, E.J., Loeb, L.A.: Mutational heterogeneity in human cancers: origin and consequences. Annu. Rev. Pathol. **5**, 51–75 (2010)
5. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz Jr., L.A., Kinzler, K.W.: Cancer genome landscapes. Science **339**, 1546–1558 (2013)
6. International Cancer Genome Consortium (ICGC). https://icgc.org/
7. Zhang, K., Wang, H.: Cancer genome atlas pan-cancer analysis project. Zhongguo Fei Ai Za Zhi **18**, 219–223 (2015)
8. Cancer Genome Atlas Research, N., et al.: The cancer genome atlas pan-cancer analysis project. Nat. Genet. **45**, 1113–1120 (2013)
9. Pancancer Analysis of Whole Genomes (PCAWG). https://dcc.icgc.org/pcawg
10. The Cancer Genome Atlas (TCGA). https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
11. Consortium, I.T.P.-C.A.o.W.G.: Pan-cancer analysis of whole genomes. Nature **578**, 82–93 (2020)
12. Javadekar, S.M., Raghavan, S.C.: Snaps and mends: DNA breaks and chromosomal translocations. FEBS J. **282**, 2627–2645 (2015)
13. Li, Y., et al.: Patterns of somatic structural variation in human cancer genomes. Nature **578**, 112–121 (2020)
14. Polak, P., et al.: Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature **518**, 360–364 (2015)
15. Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M., Nik-Zainal, S.: Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. Genome Res. **28**, 1264–1271 (2018)
16. Cheloshkina, K., Poptsova, M.: Tissue-specific impact of stem-loops and quadruplexes on cancer breakpoints formation. BMC Cancer **19**, 434 (2019)